

Sunfall: A Collaborative Visual Analytics System for Astrophysics

Cecilia R. Aragon, *Member, IEEE*, Stephen J. Bailey, Sarah Poon, Karl J. Runge, and Rollin C. Thomas

Abstract—Computational and experimental sciences produce and collect ever-larger and complex datasets, often in large-scale, multi-institution projects. The inability to gain insight into complex scientific phenomena using current software tools is a bottleneck facing virtually all endeavors of science. In this paper, we introduce Sunfall, a collaborative visual analytics system developed for the Nearby Supernova Factory, an international astrophysics experiment and the largest data volume supernova search currently in operation. Sunfall utilizes novel interactive visualization and analysis techniques to facilitate deeper scientific insight into complex, noisy, high-dimensional, high-volume, time-critical data. The system combines novel image processing algorithms, statistical analysis, and machine learning with highly interactive visual interfaces to enable collaborative, user-driven scientific exploration of supernova image and spectral data. Sunfall is currently in operation at the Nearby Supernova Factory; it is the first visual analytics system in production use at a major astrophysics project.

Index Terms—Data and knowledge visualization, scientific visualization, visual analytics, visual exploration, astrophysics.

1 INTRODUCTION

Many of today's important scientific breakthroughs are being made by large, interdisciplinary collaborations of scientists working in geographically widely distributed locations, producing and collecting vast and complex datasets. These large-scale science projects require software tools that support, not only insight into complex data, but collaborative science discovery. Visual analytics approaches, combining statistical algorithms and advanced analysis techniques with highly interactive visual interfaces that support collaborative work, offer scientists the opportunity for in-depth understanding of massive, noisy, and high-dimensional data. Astrophysics in particular lends itself to a visual analytics approach due to the inherently visual nature of much astronomical data (including images and spectra).

One of the grand challenges in astrophysics today is the effort to comprehend the mysterious “dark energy,” which accounts for three-quarters of the matter/energy budget of the universe. The existence of dark energy may well require the development of new theories of physics and cosmology. Dark energy acts to accelerate the expansion of the universe (as opposed to gravity, which acts to decelerate the expansion). Our current understanding of dark energy comes primarily from the study of supernovae.

The Nearby Supernova Factory (SNfactory) [1] is an international astrophysics experiment designed to discover and measure Type Ia supernovae in greater number and detail than has ever been done before. These supernovae are stellar explosions that have a consistent maximum brightness, allowing them to be used as “standard candles” to measure distances to other galaxies and to trace the rate of expansion of the universe and how dark energy affects the structure of the cosmos. The SNfactory receives 50-80 GB of image data per night, which must be processed and examined by teams of domain experts within 12-24 hours to obtain maximum

scientific benefit from the study of these rare and short-lived stellar events.

In order to facilitate the supernova search and data analysis process and enable scientific discovery for project astrophysicists, we developed Sunfall (SuperNova Factory AssembLy Line), a collaborative visual analytics system for the Nearby Supernova Factory that has been in production use for over a year. Sunfall incorporates sophisticated astrophysics image processing algorithms, machine learning capabilities including boosted trees and support vector machines, and astronomical data analysis with a usable, highly interactive visual interface designed to facilitate collaborative decision making. An interdisciplinary group of physicists, astronomers, and computer scientists (with specialties in machine learning, visualization, and user interface design) were involved in all aspects of Sunfall design and implementation.

This paper is organized as follows. Section 2 describes the astrophysics background for supernova detection and spectral analysis, including the SNfactory project data flow. Section 3 contains information on previous systems built for other supernova experiments. Section 4 discusses the Sunfall design approach, and Section 5 describes the Sunfall architecture, including its four major components: Search, Workflow Status Monitor, Data Forklift, and Supernova Warehouse. Section 6 discusses our conclusions and lessons learned, and Section 7 presents ideas and plans for future work.

2 SCIENCE BACKGROUND

The discovery of dark energy is primarily due to observations of Type Ia supernovae at high redshift (up to $z = 1$, or a lookback time of about 8 billion years) [2, 3]. This supernova cosmology technique hinges on the ability to reliably compare luminosities of high-redshift events to those at low redshift, necessitating detailed study of low-redshift events. Large-scale digital sky surveys are now being planned in order to constrain the properties of dark energy. These studies typically involve high-volume, time-constrained processing of wide-field CCD images, in order to detect Type Ia supernovae and capture detailed images and spectra on multiple nights over their lifespans.

The task of supernova detection is extremely challenging, involving searching for very rare, short-lived events. A stellar explosion resulting in a Type Ia supernova will occur on average only once or twice per millennium in a typical galaxy of 400 billion stars, and then, will remain visible for only a few weeks to a few

-
- Cecilia R. Aragon is with the Computational Research Division, Lawrence Berkeley National Laboratory, MS 50F-1650, One Cyclotron Road, Berkeley, CA 94720. E-Mail: aragon@hpcrd.lbl.gov.
 - Stephen J. Bailey and Rollin C. Thomas are with the Physics Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720. E-Mail: sjbailey@lbl.gov, rctomas@lbl.gov.
 - Sarah Poon and Karl J. Runge are with the Space Sciences Laboratory, University of California, Berkeley, CA 94720. E-Mail: sspoon@lbl.gov, kjrunge@lbl.gov.

months. Additionally, the maximum scientific benefit is obtained if the supernova is discovered in the first two weeks before it attains peak brightness. To further add to the challenge, the imaging data from which supernovae must be detected are extremely noisy, corrupted with spurious objects such as cosmic rays, satellite tracks, asteroids, and CCD artifacts such as diffraction spikes, saturated pixels, and ghosts.

The Nearby Supernova Factory is an international project designed to tackle these ambitious goals and to accumulate the largest homogeneously calibrated spectrophotometric (containing both images and spectra) dataset of Type Ia supernovae in the “nearby” redshift range ($0.03 < z < 0.08$, or about 0.4 to 1.1 billion light years distant) ever studied [1]. On a typical night, 50-80 GB of data (approximately 30,000 images containing 600,000 potential supernovae) are received by SNfactory image-processing software. These data are processed overnight; the best candidates are selected each morning by humans for further follow-up measurements. Likely candidate supernovae are sent to a dedicated custom-built spectrograph for follow-up imaging and spectrography. The resulting spectra typically each contain 2000-3000 data points of flux as a function of wavelength. The SNfactory database is expected to be released to astronomers worldwide and become a definitive resource for measurements of dark energy. This program is the largest data volume supernova search currently in operation.

The SNfactory obtains wide-field imaging data from the Near Earth Asteroid Tracking program (NEAT) [4] using the 112 CCD QUEST-II camera of the Mt. Palomar Oschin 1.2-m telescope [5], covering 8.5 square degrees per exposure. (For comparison, the full moon covers 0.2 square degrees.)

Images are transferred from Mt. Palomar via the High Performance Wireless Research and Education Network (HPWREN) to the High Performance Storage System (HPSS) at the National Energy Research Scientific Computing Center (NERSC) in Oakland, California. Each morning, SNfactory search software running on NERSC’s 700-node computing cluster, the Parallel Distributed Systems Facility (PDSF), matches images of the same area of the sky, processes them to remove noise and CCD artifacts, then performs an image subtraction from previously observed reference images on each set of matched images (Figure 1) [6, 7].

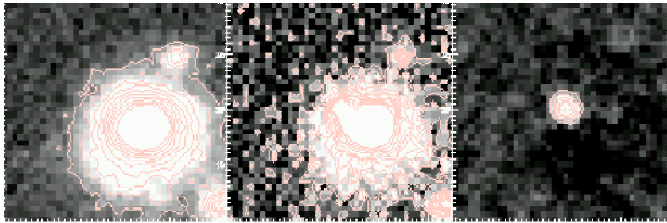


Figure 1. Example supernova images, from L to R: reference image of galaxy, new image of galaxy plus new bright spot, subtracted image reveals new supernova.

Supernova candidates are identified from among the over 600,000 objects processed per night by a set of image features including object shape (roundness and contour irregularity computed from Fourier contour descriptors) [8, 9], position, distance from nearest object, and motion. These features are used as input to machine learning algorithms that select candidates to be sent to humans for scanning and vetting. Promising supernova candidates that pass the human scanning and vetting procedure are sent for confirmation and spectrophotometric follow-up by SNIFS (the SuperNova Integral Field Spectrograph) [1] on the University of Hawaii 2.2m telescope on Mauna Kea.

Candidates are imaged through a 15 x 15 microlens array on SNIFS and the spectral data is saved to a database in France. Currently the SNfactory has collected spectrophotometric data on over 150 Type Ia supernovae, and more are being gathered each night. This dataset is expected to be an invaluable resource for

astrophysicists and cosmologists studying dark energy, the expansion rate of the universe, and supernova formation.

Spectra are the best key to understanding the actual physics of exploding stars. Understanding Type Ia supernovae through their spectra is important because it places constraints on their presupernova evolution, the immediate stellar environment of the event, and provides clues that can be turned into correlations between peak brightness and the physics of the explosion. With a better understanding of the physics of Type Ia supernovae through their spectra, it is believed that their utility as standardized candles for cosmological distance measurements can be refined for their use in future precision cosmology experiments such as SNAP, DESTINY, or ADEPT [10-12].

3 RELATED WORK

Since the discovery of dark energy via studies of Type Ia supernovae nearly a decade ago [2, 3], astrophysicists have developed several large-scale supernova searches to collect as much data as possible on these rare events. These searches typically rely on custom image subtraction software containing highly complex, hand-tuned heuristics and “cuts” to extract supernova candidates. They are often plagued by large numbers of false positives, which must then be screened out by humans in a labor-intensive process. For example, the 2005 Sloan Digital Sky Survey II (SDSS-II) supernova program generated approximately 4,000 objects per night which needed to be visually checked by humans for verification [13]. The ESSENCE and SNLS supernova searches both resulted in 100–200 objects to scan per night with a significantly smaller input data load than the SNfactory [13].

After supernova data has been collected, they are typically stored in a database accessible via a web-based interface. These web sites often present information in tabular form, listing the supernova coordinates and providing links to images. They are designed to provide all the necessary information for astronomers to observe supernovae with their own telescopes. Examples are SNLS (Supernova Legacy Survey) and SDSS (Sky Survey) [14, 15]. However, none of these groups have built a complete, automated visual analytics system that encompasses the entire data search, analysis, and scientific discovery process.

4 SUNFALL DESIGN PROCESS

In order to design an effective collaborative visual analytics system for the SNfactory, we first conducted, in mid-2005, an extensive, two-month evaluation of the existing SNfactory procedures and environment. Data sources used for evaluation included individual interviews, observation of team members performing typical project tasks, review of existing source code, literature reviews, examination of other supernova search projects, and consultation with physicists and computer scientists with relevant experience building similar scientific software systems.

We conducted over 100 hours of interviews with LBNL scientists, postdoctoral researchers, and students, and SNfactory collaboration members outside LBNL. This included 19 current and former team and collaboration members, and 12 scientists with relevant experience outside the SNfactory collaboration. We also performed a detailed software review of over 150,000 lines of existing SNfactory legacy code in C++, IDL, Perl, and shell scripts.

We established the following requirements for the Sunfall software framework: It must encompass the entire scientific data capture, processing, storage, and analysis process, enable collaborative, time-critical scientific discovery, and incorporate SNfactory legacy code (custom astronomical image-processing algorithms). It needed to reduce the number of false positives sent to humans and improve the quality of the subtraction image processing. It must automate repetitive data transfers and other manual tasks to leverage domain experts’ unique processing and image recognition skills to the maximum extent.

The Sunfall user interface was designed and implemented using participatory and iterative design techniques; for example, interactive prototypes were used to evaluate areas where existing interfaces did not support scientists' workflow. Scientists' feedback sometimes led to major redesigns of the interface. The system was implemented from the beginning with this possibility in mind, so that changes were easily made and accepted as an appropriate part of the development process [16].

Current and previous versions of Sunfall have been in operation at the SNfactory for over a year.

5 SUNFALL ARCHITECTURE & COMPONENTS

Sunfall contains four major components: Search, Workflow Status Monitor, Data Forklift, and Supernova Warehouse (Figure 2). This section will describe each component's design and structure in detail.

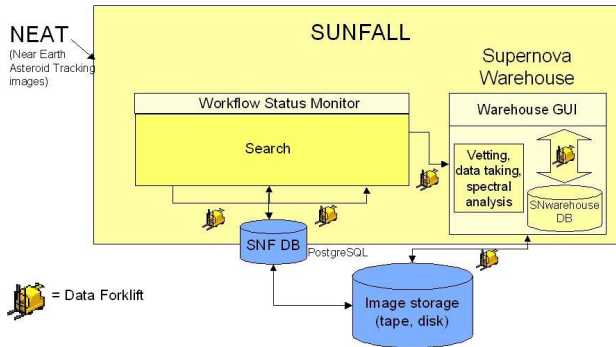


Figure 2. Sunfall architecture diagram, depicting the four components (Search, Workflow Status Monitor, Data Forklift, and Supernova Warehouse) and data flow between the components

Search incorporates SNfactory legacy software for starfield image processing and subtraction, and includes machine learning algorithms and novel Fourier contour descriptor algorithms to reduce the number of false positive supernova candidates. The Workflow Status Monitor is a web-based monitor to facilitate collaboration and improve project scientists' situational awareness of the data flow by displaying all relevant workflow (search pipeline) status data on a single site.

The Data Forklift is a middleware mechanism consisting of a coordinator and a suite of services to automate astronomical data transfers in a secure, reliable, extensible, and fault-tolerant manner. The Data Forklift also provides the middleware for the other three components, transferring data between heterogeneous systems, databases and formats securely and reliably.

The Supernova Warehouse (SNwarehouse) is a comprehensive supernova data management, workflow visualization, and collaborative scientific analysis tool. The SNwarehouse contains a PostgreSQL database, middleware consisting of Forklift mechanisms, and a graphical user interface implemented in Java.

5.1 Search

The supernova search component of Sunfall is an example of *analytic discourse*, defined in *Illuminating the Path* as “the interactive, computer-mediated process of applying human judgment to assess an issue,” [17] applied to the realm of astrophysics. By facilitating scientific analytic discourse, Sunfall Search has increased efficiency of supernova detection and enabled more effective human intervention.

The core supernova search software runs each night on NERSC's 700-node computing cluster, matches images of the same area of the

sky, processes them to remove noise and CCD artifacts, then performs an image subtraction from previously observed reference images on each set of matched images (Figure 1). This software is implemented in a variety of languages, including C++, IDL, C, Perl, and shell scripts (due to legacy requirements).

The legacy search software computed approximately 20 photometric and geometric features on each of over 600,000 supernova candidate subimages per night and applied threshold “cuts” to each of these features. Subimages that satisfied all thresholds were sent to human scanners who selected potential supernova candidates for follow-up study. The majority of subimages that passed the thresholds were false positives that humans had to manually reject. These manually tuned thresholds were brittle and often resulted in both missed supernovae and too many false positives for human scanners to evaluate daily.

Machine learning algorithms, including support vector machines and boosted trees, were incorporated into the search software. By replacing simple threshold “cuts” on the image features with classifiers of the high-dimensional data in the feature space, we reduced the number of false positives by a factor of 10 while increasing our rate of supernova detection. We applied several different classification algorithms, including boosted trees, random forests, support vector machines, and combinations of the above. All classifiers provided significantly better performance over the conventional threshold cuts. Boosted trees produced the best classification performance overall (Figure 3) and significantly decreased scientists' workload by reducing the number of false positives. This resulted in a labor savings of nearly 90%, where the process of scanning went from taking 6-8 people several hours a day to taking one person a couple of hours each day. This process is described in more detail in [6, 13].

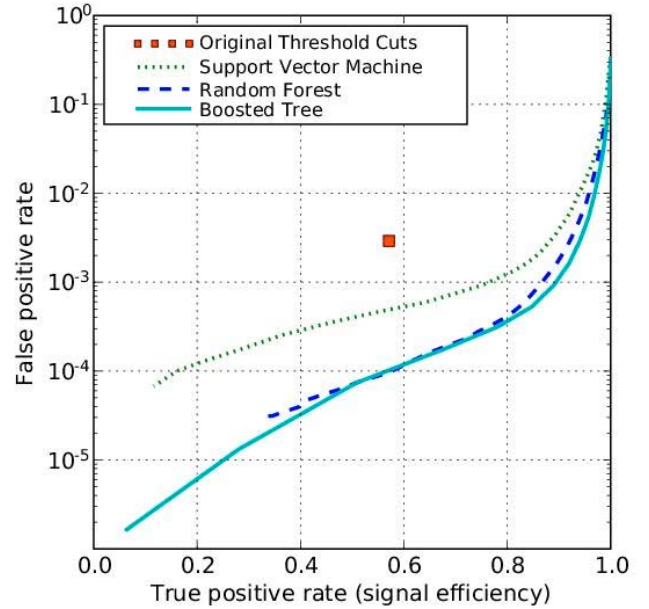


Figure 3. Comparison of Boosted Trees (cyan solid line), Random Forest (blue dashed line), and SVM (green dotted line) for false positive identification fraction vs. true positive identification fraction. The red square shows the performance of the original threshold cuts. The lower right corner of the plot represents ideal performance. (from S. Bailey et al.)

For confirmation and selection of promising supernova candidates, the scanners use a visual interface (Figure 4) to view and analyze the candidates. This visual scanning interface was redesigned to increase interactivity and allow humans to focus exclusively on the imagery itself. Previously, operators of the interface spent much of their time cutting and pasting long

filenames, pausing the scanning process to look up data on several different external web sites, and waiting for image data to be retrieved from tape storage. By applying intelligent algorithms to the search data itself, we were able to leverage the unique human abilities that enable the final detection of true supernova candidates, and minimize not only the number of false positives, but also increase the efficiency of the spectroscopic follow-up.

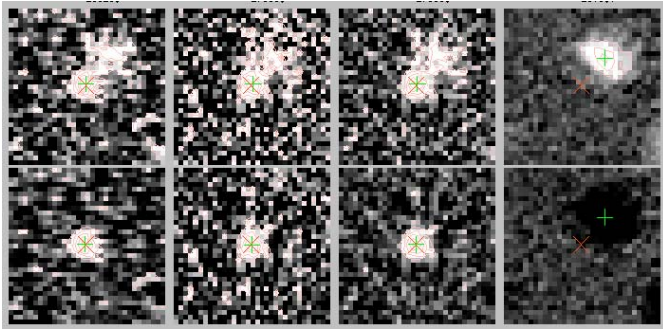


Figure 4. Visual scanning interface. Images L to R, top row: 3 new images, reference image. Bottom row: 3 subtractions, negative subtraction.

5.2 Workflow Status Monitor

The Workflow Status Monitor is a web-based monitor to improve project scientists' situational awareness of the data flow by synthesizing diverse information flows from various systems and displaying critical workflow status data on a single site (Figure 5).

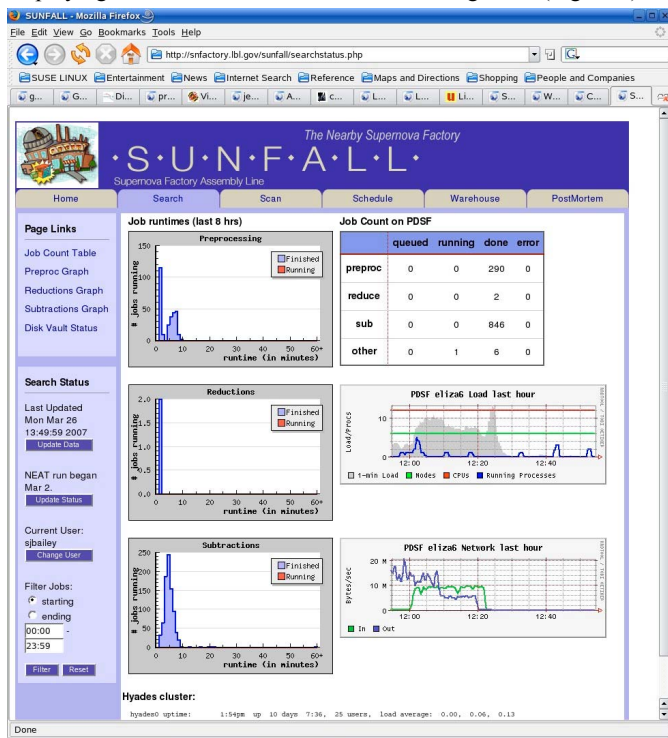


Figure 5. Sunfall Workflow Status Monitor

The supernova search software is highly parallelized as 30,000 images are queued for processing in a multi-stage pipeline that runs on NERSC's 700-node computing cluster, the Parallel Distributed Systems Facility (PDSF). Nodes frequently go down or jobs fail, and failures must be detected promptly and jobs resubmitted quickly due to the time-critical nature of the search. Detection of such failures was challenging and time-consuming before the deployment of the Workflow Status Monitor. The monitor displays graphs and

visual displays of job completion times on PDSF, job queues, PDSF node uptimes, and disk vault loads. The interface was prototyped, evaluated for efficacy by project scientists, and implemented via PHP on a web site hosted on an SNfactory server running SuSE Linux 9.3.

The status monitor also displays date and time information at project locations in several formats relevant to astronomers. One author's previous experience with a status monitor for the Mars Exploration Rovers [18] demonstrated that apparently minor features, such as a clock that displayed the time on Mars and at various project locations on Earth, yielded tremendous improvements in scientists' situational awareness and may very well have prevented critical errors.

Additional information displayed includes telescope scheduling, supernova candidates found, and current operational status of telescopes, spectrographs, and cameras used in processing SNfactory data.

5.3 Data Forklift

Each night, over 50 GB of heterogeneous, distributed supernova data arrives at the SNfactory and needs to be processed, managed, analyzed, queried, and displayed – all within a time period of 24 hours or less (ideally 12 hours). The data is distributed over wide geographic distances; some of it is hosted on systems outside the SNfactory's control, and some on unreliable systems with frequent downtime.

To solve these data management problems, we designed and implemented the Sunfall Data Forklift, a suite of services for automated data retrieval, storage, movement, querying, and staging for display. The Data Forklift consists of a coordinator and a set of individual services, each customized for its particular task, but having key attributes in common.

The Data Forklift supports "different place, asynchronous" collaborative scientific work by facilitating data and information transfer amongst a geographically separated team. Due to the time-critical nature of the data collection (telescopes must be operated at night and are located in different time zones), tasks must take place at distinct, specified times.

All Data Forklift services share the following properties:

The mechanism resides at a central location (an SNfactory server), but processes data remotely from widely separate geographic locations, including making connections to external machines not under the SNfactory's control.

All data transfers are secure, utilizing encrypted channels and authentication.

The Forklift services and coordinator are reliable and self-restarting. The Forklift coordinator restarts itself if its process dies or if the server is rebooted.

The mechanism is fault-tolerant; services detect unreliable connections and recover from errors in data transfer. Partially transferred files are detected and retransferred, and the process restarts the transfer at the point of failure.

The mechanism is extensible (new tasks can easily be added). The Forklift coordinator was designed so that new services can easily be added to a central table.

The Data Forklift enables data retrieval from external sources. Many astronomical databases are web-based, and designed for interactive rather than automated retrieval. Forklift services handle such interactions in a fault-tolerant manner. SNfactory spectral data is stored at a central processing database in France, and must be retrieved according to external requirements.

User actions can trigger Data Forklift requests. For example, whenever a supernova candidate is saved to the Supernova Warehouse (SNwarehouse) database, several Forklift services are automatically started, retrieving related image

data from tape storage, accessing external web-based asteroid, galaxy, and other astronomical databases, converting image files into display format, and staging the data for display.

Forklift services act as middleware for the Supernova Warehouse, retrieving data from internal and external databases, performing program logic, and staging data for display by the SNwarehouse GUI.

The Data Forklift is written in Perl, and runs under SuSE Linux 9.3 on an SNfactory server.

5.4 Supernova Warehouse

The Supernova Warehouse (SNwarehouse) is a comprehensive supernova data management, workflow visualization, and collaborative scientific analysis application. It consists of a PostgreSQL database hosted on a dedicated SNfactory database server running SuSE Linux 9.3, Data Forklift services as middleware, and a graphical user interface (GUI) written in Java. SNwarehouse supports collaborative remote asynchronous work in several different ways.

Collaboration members can access the GUI from any networked computer worldwide via a Forklift remote deployment mechanism. Security is provided via password authentication and encrypted communication channels. SNwarehouse furnishes project scientists with a shared workspace that enables easy distribution, analysis, and access of data. Collaboration members can view, modify, and annotate supernova data, add comments, change a candidate's state, and schedule follow-up observations from work, home, while observing at the telescope, or when attending conferences. This access is critical due to the 24/7 nature of SNfactory operations. All transactions are recorded in the SNwarehouse database, and the change history and provenance of the data is permanently stored (records cannot be deleted in order to maintain the change history) and continuously visible to all authenticated users.

SNwarehouse centralizes data from multiple sources and supports task-oriented workflow. Project members perform well-defined tasks, such as vetting, scheduling, and analyzing targets, which collectively accomplish the goal of finding and following type Ia supernovae. Typically, an individual or small group performs a given task, and the results of the task provide inputs for the next task in the workflow, often performed by another set of group members. Thus, the inputs and outputs of any task must be well-defined and easily recognizable.

5.4.1 SNwarehouse Overview

SNwarehouse's interaction design takes the approach of overview, filter and drill down to details. The main overview page (Figure 6) displays two tightly coupled representations of the list of targets registered in the database. The top visualization plots the targets in the sky; below is a sortable, tabular representation of the same data. From here, drill down depends on the defined tasks.

5.4.2 Supernova Candidate Vetting

"Vetting" involves a process of scientific analytic discourse where the scientist must quickly decide, based on limited data, how to allocate scarce and expensive telescope time to the night's supernova candidates. At this point, before spectra are taken, it is often unknown whether a target candidate is a supernova, so the scientist must gather as much relevant data as rapidly as possible to make an informed prediction. This task involves retrieving any available images of the target's location in the sky prior to discovery and querying several external astronomical databases.

Once a target is saved as a candidate supernova, the target name will appear in the SNwarehouse overview table color-coded in red, indicating that the target is newly discovered. On a rolling basis, a background process checks if previous data products exist on disk. If found, these data products are registered with the database, and a "!" appears next to the target name, a flag signaling newly found images.

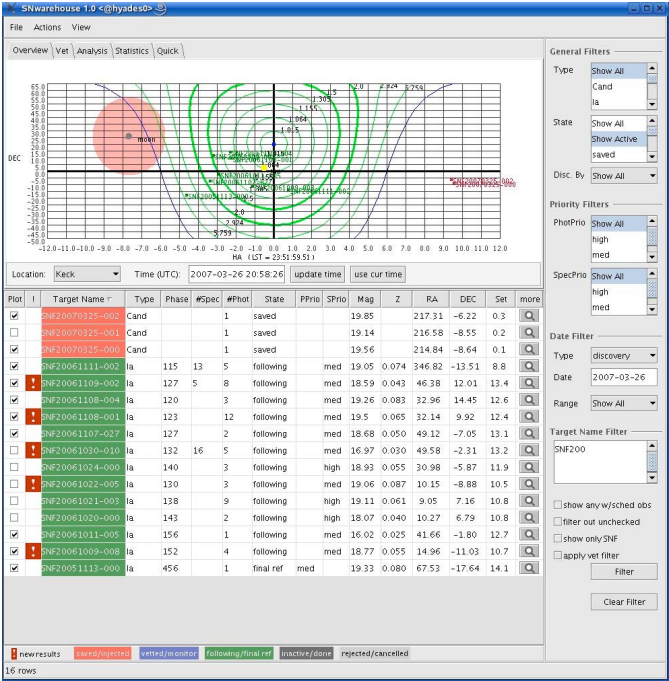


Figure 6. SNwarehouse overview tab

The combination of these two visual clues allows the scientist to quickly see which targets need evaluation. Once this determination is made, the vetter will then open the Details View for the target (Figure 7).

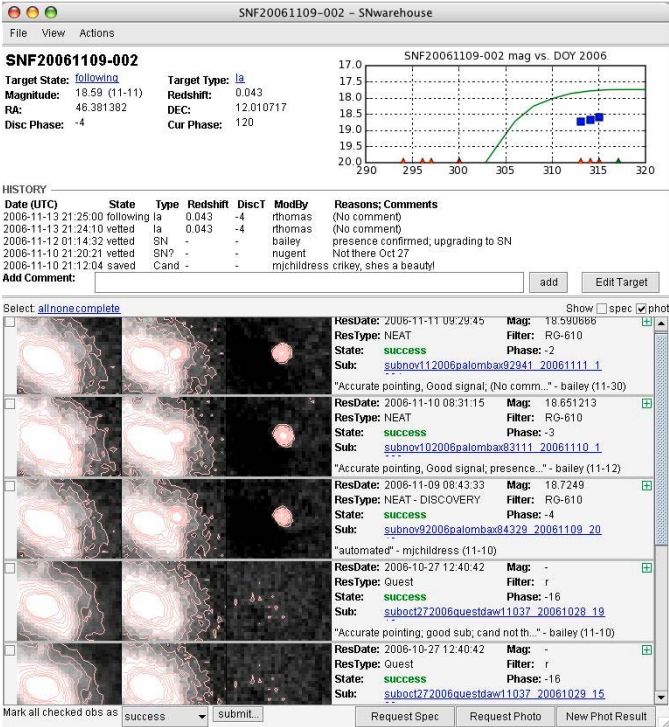


Figure 7. SNwarehouse Details View

At this stage, the vetter is trying to determine whether this target is potentially a type Ia supernova, based on how consistently the magnitude is rising in comparison to standard Ia's. The Details View helps the vetter make this determination in two ways. First, a visualization of the magnitudes of all target observations against a standard Ia lightcurve (Figure 8) allows the vetter to quickly

determine if the magnitudes are rising like those of a typical Ia. In addition, the vetter can compare the discovery image with prior images taken at this particular set of coordinates in the sky (Figure 9).

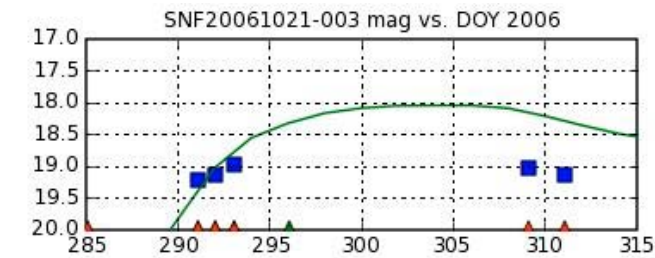


Figure 8. Supernova lightcurve. Here we see how the rise in magnitude of this target (blue squares) compares to the lightcurve of a standard SN Ia (green curve).

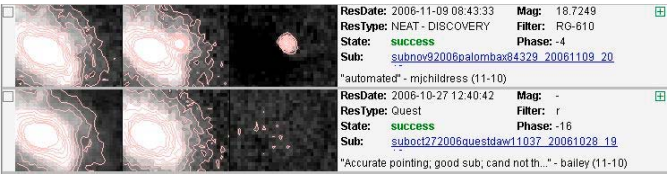


Figure 9. The target is visible at the time of discovery but not visible on a prior date.

5.4.3 Observation Scheduling

Prior to followup observation at the telescope, a schedule is made based on the vetter’s assessments to order and allocate telescope time. In SNwarehouse, the scheduler starts by filtering for only those targets with observation requests. With this list, the scheduler must order and assign exposure times for each target using a variety of techniques. Exposure times are automatically determined using a lookup table based on the phase and redshift of the target. Considering these exposure times, the scheduler must also order the target list according to when the target will be visible in the sky by the telescope. The Sky visualization offers insight for this task (Figure 10).

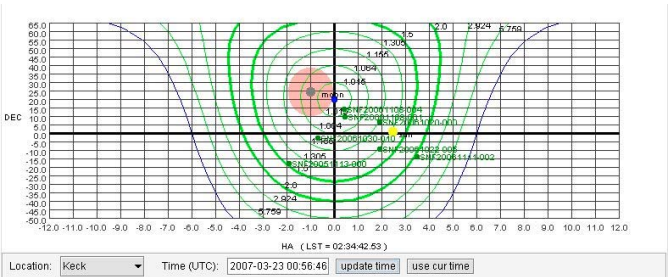


Figure 10. The Sky visualization plots declination vs. hour angle (astronomical sky coordinates) in a view preferred by domain experts. The green lines represent airmass (the thickness of the atmosphere) at particular coordinates at the specified time (original design and implementation by G. Aldering and R. Quimby).

The Sky visualization depicts the positions of targets in the sky at a given time and ground location. The green lines represent airmass (the thickness of the atmosphere) for target coordinates at the specified time. The blue line represents astronomical twilight, and the red circle around the moon is the “moon exclusion zone,” the area where light cast by the moon makes it impossible to view a target. The yellow circle is the sun. Major telescope names and corresponding latitudes and longitudes are displayed on a drop-down menu so the visualization can be used worldwide. The time can be changed so the viewer can plan observations for the remainder of the

night. If the target appears within the blue twilight line, it is visible to the telescope, assuming good weather. The scheduler must note the rise and set times of each target when creating the schedule for the night.

5.4.4 Data Taking

During each night of spectral observations with SNIFS at the University of Hawaii 2.2m telescope, an observer needs to point the telescope at each scheduled target in order. In the event of weather or mechanical problems, a rescheduling decision must be made quickly and efficiently. A visual data taking interface facilitates this process. Much of the observation procedure is automated, allowing the observer to concentrate on troubleshooting, rescheduling, and determining the success of each target observation (Figure 11).

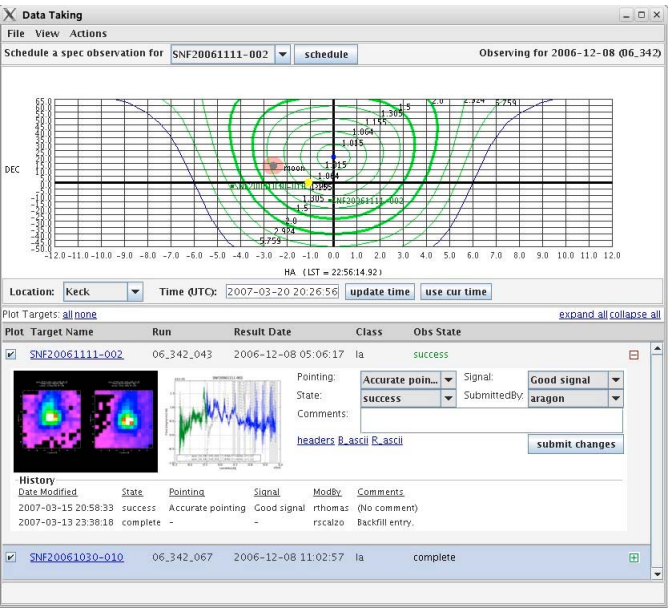


Figure 11. SNwarehouse Data Taking window. The observer can follow the targets on the Sky visualization, take notes on the success or failure of each observation, telescope status and weather conditions, and reschedule targets if necessary.

5.4.5 Analysis and Post-Mortem

Once observation at the telescope is complete, the data products (images and spectra) from the telescope are registered with the database and marked with a “!” in the SNwarehouse overview page. The next task is spectral processing and analysis, known as the “post-mortem” process. The scientist performing post-mortem filters for targets with new data products that are being followed (highlighted in green and blue, depending on the type of followup) and opens up the details view to evaluate the telescope results (Figure 12).

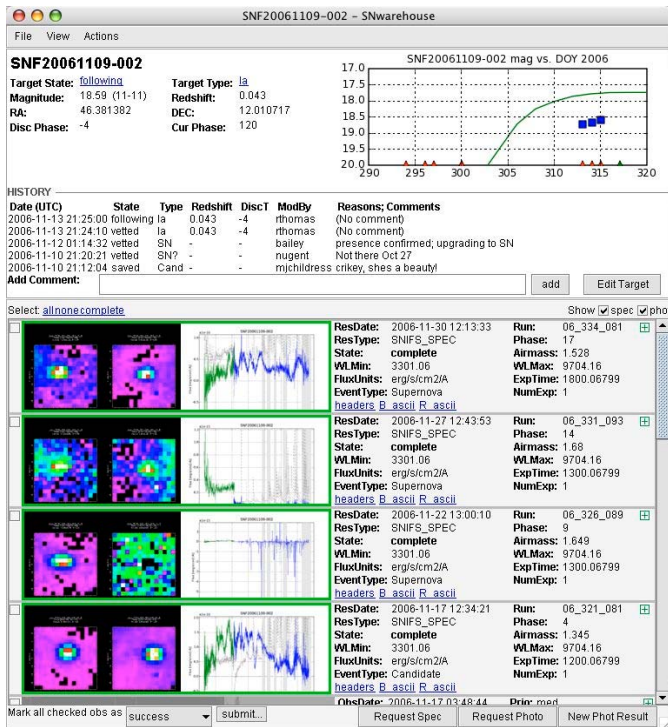


Figure 12 SNwarehouse details view depicting spectral data observations.

The raw data from the SNIFS spectrograph is complex and requires a significant amount of processing in order to yield meaning to the scientist. A visual depiction of the accuracy of the pointing and signal strength provides much more information more quickly than tables of numeric data. The two small images on the left-hand side of each observation subwindow are a custom visualization designed to indicate whether the telescope accurately pointed at the target and if a good signal was received by the spectrograph (Figure 13).

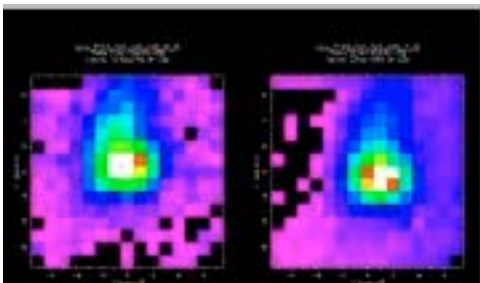


Figure 13. A successful spectral observation. The bright spot in the center represents the supernova, and indicates that the target was centered by the spectrograph and the signal was good.

Color-coding and position indicate the accuracy and signal strength of the received data. If there is a bright dot in the center of both squares, that is a clear indicator of a “good pointing.” A failed observation contains no clear bright circle in the center of the image (Figure 14). A marginal observation may contain a blue halo, indicating that too much background noise from the supernova’s host galaxy is present, and that the spectral data may be skewed (Figure 15).

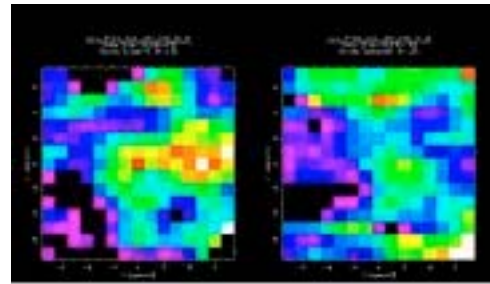


Figure 14. A failed observation. No obvious target appears visible, so no useful spectral data was captured.

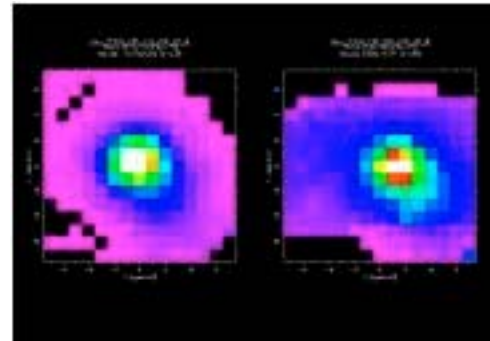


Figure 15. A marginal observation. In the right-hand image, a blue halo surrounds the bright dot in the center, indicating that too much noise from the supernova’s host galaxy has been picked up, skewing the spectral results.

Supernova scientists, like many domain experts, demonstrate strong visual pattern recognition ability in their field of expertise. They can take a single glance at a picture of a complex spectrum, and instantly determine its type, age in days before or after peak brightness, and whether it exhibits any unusual properties. An early Type Ia supernova (captured well before peak brightness) will display a certain pattern in its spectral plot (Figure 16). A later supernova (imaged a few days past peak brightness) will show other characteristic spectral lines and features (Figure 17). The spectral data is displayed in optimal form in order to facilitate domain experts’ visual pattern recognition ability. Spectral data are plotted in green and blue. The spiky grey lines depict the spectrum of the background sky. The broad grey bands represent areas of atmospheric absorption. Due to the complexity of the data (thousands of points of flux vs. wavelength for each observation) and the necessity to make rapid, accurate decisions in order to maximize the use of limited, expensive telescope time, visualization provides the most efficient solution to the problem. The scientist can also access the raw numeric data by clicking on any of the images or spectra.

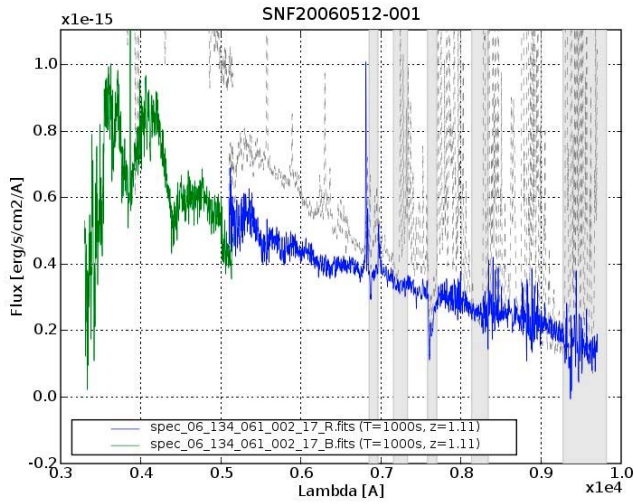


Figure 16. A spectrum from a very early Type Ia supernova, 15 days before peak brightness. The green and blue plots indicate the spectral data. The spiky grey lines in the background depict the background sky spectrum. This supernova is just starting to show the features that define a Type Ia.

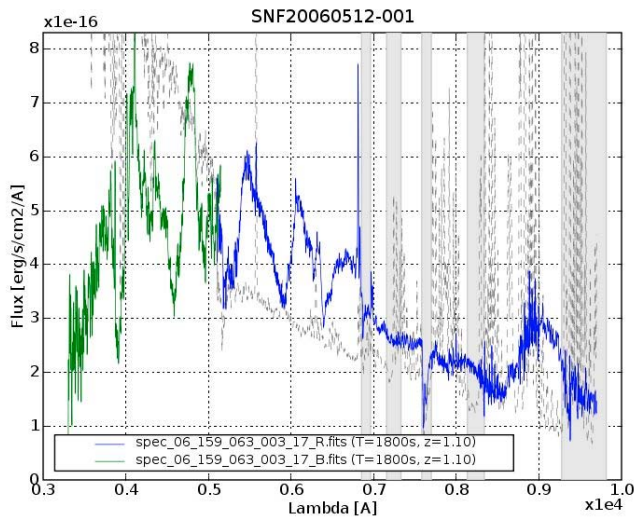


Figure 17. Spectrum of the same Type Ia supernova at 10 days after peak brightness. Note the different profile of the spectrum. To a domain expert, each of the peaks and troughs has meaning.

6 CONCLUSION

Sunfall, a collaborative visual analytics system in operation at a large-scale astrophysics project, has demonstrated that such systems that facilitate scientific analytic discourse and computer-supported collaborative work can have a positive impact on data-intensive science. In the process of design and implementation, we learned that an interdisciplinary team incorporating specialists from several fields, including scientific domain experts, will be most effective in designing an effective visual analytics system for science.

7 FUTURE WORK

Current plans for improvements to Sunfall include the development of a multi-dimensional visual analytics tool for supernova spectra classification. We are currently applying several different clustering algorithms to the normalized spectral data. Tools are under

development to apply various transformations to the calibrated and normalized spectra. The end goal is feature detection and similarity detection across supernova spectra. The optimal presentation of these clusters is still an open problem. We hope these visual analytic tools will facilitate the search for new correlations between classes of supernovae.

ACKNOWLEDGMENTS

We would like to thank the scientists of the SNfactory collaboration for their time and detailed feedback. This work was supported in part by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098, and by the Director, Office of Science, Office of High Energy Physics, of the U.S. Department of Energy under Contract No. DE-FG02-92ER40704, and by a grant from the Gordon & Betty Moore Foundation. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

- [1] G. Aldering, G. Adam, P. Antilogus, P. Astier, R. Bacon, et al., "Overview of the Nearby Supernova Factory," *Proceedings of the SPIE*, vol. 4836, 2002.
- [2] S. Perlmutter, G. Aldering, G. Goldhaber, et al., "Measurements of Omega and Lambda from 42 High-Redshift Supernovae," *Astrophysical Journal*, vol. 1999, pp. 565-586, 1999.
- [3] A. G. Riess, A. V. Filippenko, et al., "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant," *Astrophysical Journal*, vol. 1998, pp. 1009-1038, 1998.
- [4] NEAT, Near Earth Asteroid Tracking, <http://neat.jpl.nasa.gov>, 2007.
- [5] QUEST, Palomar-QUEST Survey, <http://hepwww.physics.yale.edu/quest/palomar.html>, 2002.
- [6] R. Romano, C. Aragon, and C. Ding, "Supernova Recognition Using Support Vector Machines," *Proceedings of the 5th International Conference of Machine Learning Applications*, Orlando, FL, 2006.
- [7] W. M. Wood-Vasey, "Rates and Progenitors of Type Ia Supernovae," Ph.D. dissertation, University of California, Berkeley, 2004.
- [8] C. Aragon and D. B. Aragon, "A Fast Contour Descriptor Algorithm for Supernova Image Classification," *Proc. SPIE Symposium on Electronic Imaging: Real-Time Image Processing*, San Jose, CA, 2007.
- [9] C. T. Zahn and R. Z. Roskies, "Fourier descriptors for plane closed curves," *IEEE Trans. Computers*, vol. 21, pp. 269-281, 1972.
- [10] D. Benford and T. Lauer, "Destiny: A Candidate Architecture for the Joint Dark Energy Mission," *Space Telescopes and Instrumentation I: Optical, Infrared, and Millimeter*, *Proc. of SPIE*, 2006.
- [11] SNAP, Supernova Acceleration Probe, <http://snap.lbl.gov>, 2007.
- [12] ADEPT, Advanced Dark Energy Physics Telescope, <http://www.physorg.com/news73758591.html>, 2006.
- [13] S. Bailey, C. Aragon, R. Romano, R. C. Thomas, B. A. Weaver, and D. Wong, "How to Find More Supernovae with Less Work: Object Classification Techniques for Difference Imaging," *Astrophysical Journal*, submitted for publication, 2007.
- [14] SNLS, SuperNova Legacy Survey, <http://www.cfht.hawaii.edu/SNLS/>, 2007.
- [15] SDSS, Sloan Digital Sky Survey, <http://www.sdss.org>, 2007.
- [16] C. Aragon and S. Poon, "The Impact of Usability on Supernova Discovery," Workshop on Increasing the Impact of Usability Work in Software Development, *CHI 2007: ACM Conference on Human Factors in Computing Systems*, San Jose, CA, 2007.
- [17] J. J. Thomas and K. A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*: National Visualization and Analytics Center, 2005.
- [18] R. Mak, J. Walton, L. Keely, D. Heher, and L. Chan, "Reliable Service-Oriented Architecture for NASA's Mars Exploration Rover Mission," *Aerospace, 2005 IEEE Conference*, 2005.